Long-term dependency slow feature analysis for dynamic process monitoring

Xinrui Gao*, Yuri A.W. Shardt*

** Technical University of Ilmenau, Ilmenau, Thuringia, Germany, 98684*
(e-mail: {xinrui.gao,yuri.shardt}@tu-ilmenau.de).

Abstract: Industrial processes are large scale, highly complex systems. The complex flow of mass and energy, as well as the compensation effects of closed-loop control systems, cause significance cross-correlation and autocorrelation between process variables. To operate the process systems stably and efficiently, it is crucial to uncover the inherent characteristics of both variance structure and dynamic relationship. Long-term dependency slow feature analysis (LTSFA) is proposed in this paper to overcome the Markov assumption of the original slow feature analysis to understand the long-term dynamics of processes, based on which a monitoring procedure is designed. A simulation example and the Tennessee Eastman process benchmark are studied to show the performance of LTSFA. The proposed method can better extract the system dynamics and monitor the process variations using fewer slow features.

Keywords: **Process monitoring, slow feature analysis, fault diagnosis, latent variable model**

1. INTRODUCTION

Modern industrial processes are large scale, highly complex systems with many units and equipment. A large number of measurements are set to monitor process operation, and much process data is automatically collected. To operate the process systems stably and efficiently, developing process models based on process data to describe the process characteristics is crucial. The process data has two distinctive characteristics. The first is the serious collinearity caused by mass and energy balances and the compensation effects of the closed-loop control systems. Secondly, the process system always shows significant time-dependent characteristics. For instance, a transition process would occur while state switching due to the long settling time. Similarly, oscillations would occur after a disturbance introduced because of recycle streams, heat integrations, and other connections of materials, energy, and information (Shardt et al., 2012).

Multivariate statistical process monitoring (MSPM) approaches have been extensively studied to deal with the above two problems (Qin, 2012). The most widely used method is principal component analysis (PCA) (Gajjar, Kulahci, & Palazoglu, 2018; Wise, Ricker, Veltkamp, & Kowalski, 1990), which extracts uncorrelated lower dimensional latent variables to concisely describe the main variance structure of the process observation space. Independent component analysis (ICA) recovers the independent latent variables from complex process data by leveraging high order moment information (Kano, Tanaka, Hasebe, Hashimoto, & Ohno, 2003). For the description of time-dependent characteristics, dynamic PCA (DPCA) performs the standard PCA on the augmented data with time lags to extract process dynamics (Ku, Storer, & Georgakis, 1995). However, it fails to decouple the dynamics and static variance information. Slow feature analysis (SFA) is used to concurrently monitor the process by describing the stationary distribution and dynamic behaviours separately (Shang et al., 2015). However, assuming the standard Markov property, namely that the one-step time dependency is sufficient, limits the performance of the standard SFA and has to be improved in the same way as DPCA in practice.

Thus, this paper proposes a new long-term dependency slow feature analysis (LTSFA) method that can extract longer time dependencies for the design of a process monitoring method. This approach is validated using a simulated example and the Tennessee Eastman process.

2. LINEAR SLOW FEATURE ANALYSIS

Given an m-dimensional ergodic observation signal $\boldsymbol{x}(t) = [x_1(t) \quad \cdots \quad x_m(t)]$, SFA seeks to find a latent signal $\boldsymbol{s}(t) = [s_1(t) \quad \cdots \quad s_k(t)]^\mathsf{T}$ with the slowest variation, also named slow features (SFs), to describe its time varying characteristics (Wiskott & Sejnowski, 2002), that is,

$$\begin{aligned} &\min_{g_j(\bullet)} \left\langle \dot{s}_j^2(t) \right\rangle \\ \text{s.t.} \quad &1)\ \left\langle s_j(t) \right\rangle = 0, \\ &2)\ \left\langle s_j{}^2(t) \right\rangle = 1, \\ &3)\ \forall i \neq j,\ \left\langle s_i(t) s_j(t) \right\rangle = 0 \end{aligned} \tag{1}$$

where $g_j(\bullet)$ is the input-output function that needs to be found, $\langle \bullet \rangle$ is the temporal averaging operator, and $\dot{s}(t)$ is the first-order derivative or difference with respect to time and can be approximated for discrete time series as

$$\dot{s}(t) \approx s(t) - s(t-1) \tag{2}$$

Objective (1) uses the average squared temporal difference to define the signal slowness. Constraints 1) and 2) simplify the problem without loss of generality, and constraint

2) avoids the trivial solution s_j = constant. Constraint 3) guarantees that the SFs are statistically uncorrelated so that they carry different information. It also implicitly implies that the extracted SFs are sorted by their slowness. In the linear case, the input-output function set $\boldsymbol{g} = [g_j(\cdot)]_{j=1}^{k}$ is a linear transformation

$$\boldsymbol{s}(t) = \boldsymbol{g}(\boldsymbol{x}(t)) = \mathbf{P}\boldsymbol{x}(t) \tag{3}$$

where $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_k]^\top$ is the parameter matrix.

For a mean centred signal $\boldsymbol{x}(t)$, due to

$$\begin{aligned} \langle \dot{s}_j^2(t) \rangle &= \langle (s_j(t))^2 + (s_j(t-1))^2 - 2(s_j(t)s_j(t-1)) \rangle \\ &= 2 - 2\langle s_j(t)s_j(t-1) \rangle \end{aligned} \tag{4}$$

the objective function of SFA is equivalent to the following formulation with the same constraints, that is,

$$\max_{\boldsymbol{p}} \langle s_j(t)s_j(t-1) \rangle \tag{5}$$

Using Lagrange multipliers, this optimization problem can be translated into the generalized eigenvalue decomposition (GED) problem, that is

$$\begin{aligned} &C^x(1)\mathbf{P} = C^x(0)\mathbf{P}\boldsymbol{\Lambda} \\ &\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_k\}, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \end{aligned} \tag{6}$$

where λ_j is the generalized eigenvalue of matrix pair $(C^x(1), C^x(0))$, and $C^x(\tau)$ is the symmetric version of the τ-step autocovariance matrix of $\boldsymbol{x}(t)$ defined as

$$C^x(\tau) \equiv \frac{1}{2}\left(\langle \boldsymbol{x}(t)\boldsymbol{x}^\top(t-\tau) \rangle + \langle \boldsymbol{x}(t-\tau)\boldsymbol{x}^\top(t) \rangle \right) \tag{7}$$

3. LONG-TERM DEPENDENCY SFA

3.1. Optimization problem for long-term dependency SFA

From Equation (4), it can be concluded that minimizing the average squared temporal difference of SF is equivalent to maximizing its one-step lagged autocorrelation. However, in practical systems, one-step lagged autocorrelation is far from sufficient to describe the dynamics, especially manufacturing processes with typical long-term dependency. Hence, it is necessary to improve SFA to possess long-term dependency modelling ability.

From a predictability perspective, the long-term dependency can be represented as

$$\hat{s}(t) = \beta_1 s(t-1) + \beta_2 s(t-2) + \cdots + \beta_l s(t-l) \tag{8}$$

where l is the maximum time delay. Thus, the objective of the long-term dependency SFA can be formulated as

$$\begin{aligned} \langle s(t)\hat{s}(t) \rangle &= \frac{1}{2}\langle s(t)\hat{s}(t) + \hat{s}(t)s(t) \rangle \\ &= \sum_{\tau=1}^{l} \beta_\tau C^s(\tau) = \boldsymbol{p}^\top \left(\sum_{\tau=1}^{l} \beta_\tau C^x(\tau) \right) \boldsymbol{p} \end{aligned} \tag{9}$$

where $C^s(\tau)$ and $C^x(\tau)$ are defined as in Equation (7). Let $\mathbf{X} = [\boldsymbol{x}(1) \quad \cdots \quad \boldsymbol{x}(N+l)]$ be the total sample collection of $\boldsymbol{x}(t)$. Construct the following matrices

$$\mathbf{X}_i = [\boldsymbol{x}(i) \quad \cdots \quad \boldsymbol{x}(i+N-1)] \text{ for } i = 1, \cdots, l+1 \tag{10}$$

$$\mathbf{A}_\tau = [\mathbf{X}_{l+1} \quad \mathbf{X}_{l+1-\tau}], \ \mathbf{B}_\tau = [\mathbf{X}_{l+1-\tau} \quad \mathbf{X}_{l+1}]$$

where $\tau \in \{1, 2, \cdots, l\}$ is the time delay. Then, the τ-step covariance matrix can be written as

$$\begin{aligned} C^x(\tau) &= \frac{1}{2}\left(\langle \boldsymbol{x}(t)\boldsymbol{x}^\top(t-\tau) \rangle + \langle \boldsymbol{x}(t-\tau)\boldsymbol{x}^\top(t) \rangle \right) \\ &= \frac{1}{2N}\sum_{t=l+1}^{l+N} \left(\boldsymbol{x}(t)\boldsymbol{x}^\top(t-\tau) + \boldsymbol{x}(t-\tau)\boldsymbol{x}^\top(t) \right) = \frac{\mathbf{A}_\tau \mathbf{B}_\tau^\top}{2N} \end{aligned} \tag{11}$$

Substituting Equation (11) into Equation (9), gives

$$\langle s(t)\hat{s}(t) \rangle = \frac{1}{2N}\boldsymbol{p}^\top \left(\sum_{\tau=1}^{l} \beta_\tau \mathbf{A}_\tau \mathbf{B}_\tau^\top \right) \boldsymbol{p} = \frac{1}{2N}\boldsymbol{p}^\top \mathbf{A}\boldsymbol{\Lambda}_{\boldsymbol{\beta}}\mathbf{B}^\top \boldsymbol{p} \tag{12}$$

where $\mathbf{A} = [\mathbf{A}_1 \quad \mathbf{A}_2 \quad \cdots \quad \mathbf{A}_l]$, $\mathbf{B} = [\mathbf{B}_1 \quad \mathbf{B}_2 \quad \cdots \quad \mathbf{B}_l]$, $\boldsymbol{\Lambda}_{\boldsymbol{\beta}} = \mathrm{diag}\{\beta_1\mathbf{I}, \beta_2\mathbf{I}, \cdots, \beta_l\mathbf{I}\}$.

Then, the optimization objective of LTSFA can be rewritten as

$$\begin{aligned} &\max_{\boldsymbol{p}, \boldsymbol{\beta}} \ \boldsymbol{p}^\top \mathbf{A}\boldsymbol{\Lambda}_{\boldsymbol{\beta}}\mathbf{B}^\top \boldsymbol{p} \\ &\text{s.t.} \quad 1) \quad \boldsymbol{p}^\top \mathbf{X}_{l+1}\mathbf{X}_{l+1}^\top \boldsymbol{p} = 1 \\ &\qquad 2) \quad \| \boldsymbol{\beta} \| = 1 \end{aligned} \tag{13}$$

where constraint 1) guarantees that the extracted SFs have unit variances to avoid the trivial solution. It is clear that LTSFA will reduce to SFA if the time delay is set to one.

3.2. Algorithm for long-term dependency SFA

3.2.1. Analysis of the optimization formulation

Introducing Lagrange multipliers λ_p and λ_β, Equation (13) can be converted into

$$J = \boldsymbol{p}^\top \mathbf{A}\boldsymbol{\Lambda}_{\boldsymbol{\beta}}\mathbf{B}^\top \boldsymbol{p} + \lambda_p\left(1 - \boldsymbol{p}^\top \mathbf{X}_{l+1}\mathbf{X}_{l+1}^\top \boldsymbol{p}\right) + \lambda_\beta\left(1 - \boldsymbol{\beta}^\top \boldsymbol{\beta}\right) \tag{14}$$

Taking derivatives with respective to $\boldsymbol{p}$ and $\boldsymbol{\beta}$, and setting them to zero, gives

$$\frac{\delta J}{\delta \boldsymbol{p}} = 2\mathbf{A}\boldsymbol{\Lambda}_{\boldsymbol{\beta}}\mathbf{B}^\top \boldsymbol{p} - 2\lambda_p \mathbf{X}_{l+1}\mathbf{X}_{l+1}^\top \boldsymbol{p} = 0 \tag{15}$$

$$\frac{\delta J}{\delta \boldsymbol{\beta}} = (\mathbf{I} \otimes \boldsymbol{p})^\top \left([\boldsymbol{\Gamma}_1 \quad \cdots \quad \boldsymbol{\Gamma}_l]^\top \otimes \mathbf{B} \right) \mathbf{A}^\top \boldsymbol{p} - 2\lambda_\beta \boldsymbol{\beta} = 0 \tag{16}$$

where $\otimes$ is the Kronecker product, and $\boldsymbol{\Gamma}_i$ is a selection matrix:

$$\boldsymbol{\Gamma}_i = \mathrm{diag}\{\delta_1\mathbf{I}, \cdots, \delta_j\mathbf{I}, \cdots, \delta_l\mathbf{I}\}, \ \delta_j = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \tag{17}$$

Premultiplying Equations (15) and (16) by $\boldsymbol{p}^\top$ and $\boldsymbol{\beta}^\top$ respectively, the relationship $J^* = \lambda_p = 2\lambda_\beta$ can be obtained from

$$2\lambda_p = 2\lambda_p \boldsymbol{p}^\top \mathbf{X}_{l+1}\mathbf{X}_{l+1}^\top \boldsymbol{p} = 2\boldsymbol{p}^\top \mathbf{A}\boldsymbol{\Lambda}_{\boldsymbol{\beta}}\mathbf{B}^\top \boldsymbol{p} = 2J^* \tag{18}$$

$$2\lambda_\beta = 2\lambda_\beta \boldsymbol{\beta}^\top \boldsymbol{\beta} = \boldsymbol{\beta}^\top (\mathbf{I} \otimes \boldsymbol{p})^\top \left(\begin{bmatrix} \boldsymbol{\Gamma}_1 & \cdots & \boldsymbol{\Gamma}_l \end{bmatrix}^\top \otimes \mathbf{B} \right) \mathbf{A}^\top \boldsymbol{p}$$
$$= \boldsymbol{p}^\top \left(\mathbf{A} \boldsymbol{\Lambda}_{\boldsymbol{\beta}} \mathbf{B}^\top \right)^\top \boldsymbol{p} = \boldsymbol{p}^\top \mathbf{A} \boldsymbol{\Lambda}_{\boldsymbol{\beta}} \mathbf{B}^\top \boldsymbol{p} = J^* \quad (19)$$

From Equations (15) and (18), the optimal value of Equation (13) J^* is the largest generalized eigenvalue of the matrix pair

$$\left(\mathbf{A} \boldsymbol{\Lambda}_{\boldsymbol{\beta}} \mathbf{B}^\top, \mathbf{X}_{l+1} \mathbf{X}_{l+1}^\top \right) = \left(\sum_{\tau=1}^{l} \beta_\tau C^x(\tau), C^x(0) \right)$$

The optimal solution $\boldsymbol{p}^*$ is the corresponding generalized eigenvector. However, $\boldsymbol{p}$ and $\boldsymbol{\beta}$ are coupled together and thus cannot be solved using analytical methods.

3.2.2. Iterative solution for one slow feature

Let $s(t) = \boldsymbol{x}^\top(t)\boldsymbol{p}$ be the SFs at time instant *t*, then we can form the following data matrices,

$$\boldsymbol{s}_i = \begin{bmatrix} s(i) & \cdots & s(i+N-1) \end{bmatrix}^\top = \mathbf{X}_i^\top \boldsymbol{p} \text{ for } i = 1, \cdots, l+1 \quad (20)$$

$$\widehat{\mathbf{S}} = [\boldsymbol{s}_l \quad \boldsymbol{s}_{l-1} \quad \cdots \quad \boldsymbol{s}_1], \ \mathbf{Z} = [\mathbf{X}_l \quad \mathbf{X}_{l-1} \quad \cdots \quad \mathbf{X}_1]$$

where $\mathbf{X}_i$ is defined as in Equation (10). Then, Equation (15) can be rewritten as

$$\begin{aligned} \lambda_p \mathbf{X}_{l+1} \mathbf{X}_{l+1}^\top \boldsymbol{p} &= \mathbf{A} \boldsymbol{\Lambda}_{\boldsymbol{\beta}} \mathbf{B}^\top \boldsymbol{p} = \sum_{\tau=1}^{l} \left(\beta_\tau \mathbf{A}_\tau \mathbf{B}_\tau^\top \boldsymbol{p} \right) \\ &= \mathbf{X}_{l+1} \sum_{\tau=1}^{l} \left(\beta_\tau \mathbf{X}_{l+1-\tau}^\top \boldsymbol{p} \right) + \sum_{\tau=1}^{l} \left(\beta_\tau \mathbf{X}_{l+1-\tau} \right) \mathbf{X}_{l+1}^\top \boldsymbol{p} \\ &= \mathbf{X}_{l+1} \begin{bmatrix} \boldsymbol{s}_l & \cdots & \boldsymbol{s}_1 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{X}_l & \cdots & \mathbf{X}_1 \end{bmatrix} (\boldsymbol{\beta} \otimes \mathbf{I}) \boldsymbol{s}_{l+1} \\ &= \mathbf{X}_{l+1} \widehat{\mathbf{S}} \boldsymbol{\beta} + \mathbf{Z} (\boldsymbol{\beta} \otimes \mathbf{I}) \boldsymbol{s}_{l+1} \end{aligned} \quad (21)$$

Thus, we have

$$\lambda_p \boldsymbol{p} = \left(\mathbf{X}_{l+1} \mathbf{X}_{l+1}^\top \right)^\dagger \left(\mathbf{X}_{l+1} \widehat{\mathbf{S}} \boldsymbol{\beta} + \mathbf{Z} (\boldsymbol{\beta} \otimes \mathbf{I}) \boldsymbol{s}_{l+1} \right) \quad (22)$$

where $(\bullet)^\dagger$ is the Moore-Penrose inverse. Equation (16) can be rewritten as

$$\begin{aligned} 2\lambda_\beta \boldsymbol{\beta} &= (\mathbf{I} \otimes \boldsymbol{p})^\top \left(\begin{bmatrix} \boldsymbol{\Gamma}_1 & \cdots & \boldsymbol{\Gamma}_l \end{bmatrix}^\top \otimes \mathbf{B} \right) \mathbf{A}^\top \boldsymbol{p} \\ &= (\mathbf{I} \otimes \boldsymbol{p})^\top \operatorname{diag}\{\mathbf{B}_1, \cdots, \mathbf{B}_l\} \mathbf{A}^\top \boldsymbol{p} = 2\widehat{\mathbf{S}}^\top \boldsymbol{s}_{l+1} \end{aligned} \quad (23)$$

Then, the iterative solution is:

(1) Scale $\mathbf{x}$ to zero mean and unit variance.

(2) Initialize $\boldsymbol{p}$ with a random unit vector.

(3) Iteratively solve $\boldsymbol{p}$ and $\boldsymbol{\beta}$ until the objective function J converges to the optimal value, that is,

$$\boldsymbol{s} = \mathbf{X}^\top \boldsymbol{p}, \ \boldsymbol{s} := \boldsymbol{s} / \| \boldsymbol{s} \|$$

$$\boldsymbol{\beta} = \widehat{\mathbf{S}}^\top \boldsymbol{s}_{l+1}, \ \boldsymbol{\beta} := \boldsymbol{\beta} / \| \boldsymbol{\beta} \|$$

$$\boldsymbol{p} = \left(\mathbf{X}_{l+1} \mathbf{X}_{l+1}^\top \right)^\dagger \left(\mathbf{X}_{l+1} \widehat{\mathbf{S}} \boldsymbol{\beta} + \mathbf{Z} (\boldsymbol{\beta} \otimes \mathbf{I}) \boldsymbol{s}_{l+1} \right)$$

$$J = \boldsymbol{s}_{l+1}^\top \sum_{\tau=1}^{l} \beta_\tau \boldsymbol{s}_{l+1-\tau} = \boldsymbol{s}_{l+1}^\top \widehat{\mathbf{S}} \boldsymbol{\beta}$$

If the predefined stopping condition is fulfilled, the obtained $\boldsymbol{p}$ and $\boldsymbol{\beta}$ are the optimal solution corresponding to the SF.

3.2.3. Multiple slow feature extraction through matrix deflation

After the first SF is obtained, the next one can be extracted by applying Step (3) in Section 3.2.2 to the residual data excluding the previous SF information. The residual data can be obtained by matrix deflation

$$\mathbf{X} \equiv \mathbf{X} - \boldsymbol{q}\boldsymbol{s}^\top \quad (24)$$

where the loading vector $\boldsymbol{q} = \mathbf{X}\boldsymbol{s} / \boldsymbol{s}^\top \boldsymbol{s}$ is the solution of

$$\min \| \mathbf{X} - \boldsymbol{q}\boldsymbol{s}^\top \|_{\mathrm{F}} \quad (25)$$

Iteratively performing the above procedure can obtain all the k SFs in the descending order of slowness or predictability. The different SFs are orthogonal to each other, which is the same as in the standard SFA.

3.2.4. Model developed for historical slow feature extraction and prediction

After all the SFs have been extracted, the optimal projection vectors, loading vectors, and regression weights are collected as

$$\mathbf{P} = \begin{bmatrix} \boldsymbol{p}_1 & \cdots & \boldsymbol{p}_k \end{bmatrix}^\top, \mathbf{Q} = \begin{bmatrix} \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_k \end{bmatrix}^\top, \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_k \end{bmatrix}^\top.$$

Let $\mathbf{X}^{(i)}$ be the residual data matrix after the *i*-th deflation, $\mathbf{X}^{(1)} = \mathbf{X}$ and $\boldsymbol{s}^{(i)} = \left(\mathbf{X}^{(i)} \right)^\top \boldsymbol{p}_i$ be the *i*-th SF, the observation space can be divided using Equation (24) into the SF subspace and residual subspace

$$\mathbf{X} = \boldsymbol{q}_1 \left(\boldsymbol{s}^{(1)} \right)^\top + \cdots + \boldsymbol{q}_k \left(\boldsymbol{s}^{(k)} \right)^\top + \mathbf{X}^{(k+1)} = \mathbf{Q}^\top \mathbf{S}^\top + \mathbf{X}^{(k+1)} \quad (26)$$

where $\mathbf{S} = \begin{bmatrix} \boldsymbol{s}^{(1)} & \cdots & \boldsymbol{s}^{(k)} \end{bmatrix}$ is the k SFs of the historical data $\mathbf{X}$, and is also a set of the orthogonal basis of the SF subspace. Since the residuals $\mathbf{X}^{(k+1)}$ belong to the nullspace of the projection matrix $\mathbf{P}$, we have

$$\mathbf{X}^\top \mathbf{P}^\top - \mathbf{S}\mathbf{Q}\mathbf{P}^\top = \left(\mathbf{X}^{(k+1)} \right)^\top \mathbf{P}^\top = \mathbf{0} \quad (27)$$

Hence, the historical SF model is

$$\mathbf{S} = \mathbf{X}^\top \mathbf{P}^\top \left(\mathbf{Q}\mathbf{P}^\top \right)^{-1} \quad (28)$$

For an observation $\boldsymbol{x}(t)$, its projection onto SF subspace is

$$\boldsymbol{s}(t) = \left(\mathbf{P}\mathbf{Q}^\top \right)^{-1} \mathbf{P}\boldsymbol{x}(t) \quad (29)$$

Given the historical data $\overline{\mathbf{X}} = \begin{bmatrix} \boldsymbol{x}(t) & \cdots & \boldsymbol{x}(t-l+1) \end{bmatrix}$, the one-step ahead prediction model can be obtained as

$$\begin{aligned} \hat{\boldsymbol{s}}(t+1) &= \begin{bmatrix} \hat{s}^{(1)}(t+1) & \cdots & \hat{s}^{(k)}(t+1) \end{bmatrix}^\top \\ &= \begin{bmatrix} \boldsymbol{\beta}_1^\top \overline{\boldsymbol{s}}^{(1)} & \cdots & \boldsymbol{\beta}_k^\top \overline{\boldsymbol{s}}^{(k)} \end{bmatrix}^\top = \boldsymbol{\Lambda}_{\overline{\mathbf{S}}} \operatorname{vec}\left(\boldsymbol{\Phi}^\top \right) \end{aligned} \quad (30)$$

where $\overline{\boldsymbol{s}}^{(i)} = \begin{bmatrix} s^{(i)}(t) & \cdots & s^{(i)}(t-l+1) \end{bmatrix}^\top$, $\operatorname{vec}(\bullet)$ is the column vectorization operator, $\boldsymbol{\Lambda}_{\overline{\mathbf{S}}} = \operatorname{diag}\left\{ \left(\overline{\boldsymbol{s}}^{(1)} \right)^\top \quad \cdots \quad \left(\overline{\boldsymbol{s}}^{(k)} \right)^\top \right\}$ is a blocked diagonal

matrix deduced from $\bar{\mathbf{S}}$, and $\bar{\mathbf{S}}$ is the slow feature matrix of $\bar{\mathbf{X}}$, that is,

$$\bar{\mathbf{S}} = \left[\bar{\boldsymbol{s}}^{(1)} \quad \cdots \quad \bar{\boldsymbol{s}}^{(k)}\right]^{\top} = \left(\bar{\mathbf{X}}^{\top}\mathbf{P}^{\top}\left(\mathbf{Q}\mathbf{P}^{\top}\right)^{-1}\right)^{\top} \tag{31}$$

Furthermore, one-step ahead observation can be predicted by

$$\hat{\boldsymbol{x}}(t+1) = \mathbf{Q}^{\top}\hat{\boldsymbol{s}}(t+1) \tag{32}$$

In summary, the detailed procedure of long-term dependency SFA algorithm is given in Table 1.

Table 1: The algorithm for LTSFA

Input:	$\mathbf{X} = \left[\boldsymbol{x}(1) \quad \cdots \quad \boldsymbol{x}(N+l)\right]$, $\boldsymbol{x}(t) = \left[x_1(t) \quad \cdots \quad x_m(t)\right]^{\top}$
1.	Scale $\mathbf{X}$ to zero mean and unit variance
2.	Initialize $\boldsymbol{p}$ with a random unit vector
3.	Iterate the following steps until convergence to obtain the optimal parameter pair $\left(\boldsymbol{p}^*, \boldsymbol{\beta}^*\right)$: 1） $\boldsymbol{s} = \mathbf{X}^{\top}\boldsymbol{p}$, $\boldsymbol{s} := \boldsymbol{s} / \|\boldsymbol{s}\|$ 2） Form the data matrices: $\mathbf{X}_i = \left[\boldsymbol{x}(i) \quad \cdots \quad \boldsymbol{x}(i+N-1)\right]$ for $i = 1,\cdots,l+1$, $\boldsymbol{s}_i = \left[s(i) \quad \cdots \quad s(i+N-1)\right]^{\top}$ for $i = 1,\cdots,l+1$, $\hat{\mathbf{S}} = \left[\boldsymbol{s}_l \quad \cdots \quad \boldsymbol{s}_1\right]$, $\mathbf{Z} = \left[\mathbf{X}_l \quad \cdots \quad \mathbf{X}_1\right]$ 3） $\boldsymbol{\beta} = \hat{\mathbf{S}}^{\top}\boldsymbol{s}_{l+1}$, $\boldsymbol{\beta} := \boldsymbol{\beta} / \|\boldsymbol{\beta}\|$ 4） $\boldsymbol{p} = \left(\mathbf{X}_{l+1}\mathbf{X}_{l+1}^{\top}\right)^{\dagger}\left(\mathbf{X}_{l+1}\hat{\mathbf{S}}\boldsymbol{\beta} + \mathbf{Z}(\boldsymbol{\beta}\otimes\mathbf{I})\boldsymbol{s}_{l+1}\right)$ 5） $J = \boldsymbol{s}_{l+1}^{\top}\sum_{\tau=1}^{l}\beta_{\tau}\boldsymbol{s}_{l+1-\tau} = \boldsymbol{s}_{l+1}^{\top}\hat{\mathbf{S}}\boldsymbol{\beta}$
4.	Collect $\boldsymbol{p}^*$, $\boldsymbol{\beta}^*$ into $\mathbf{P} = \left[\boldsymbol{p}_1 \quad \cdots \quad \boldsymbol{p}_k\right]^{\top}$, $\boldsymbol{\Phi} = \left[\boldsymbol{\beta}_1 \quad \cdots \quad \boldsymbol{\beta}_k\right]^{\top}$
5.	Data matrix deflation. 1） $\mathbf{X} := \mathbf{X} - \boldsymbol{q}\boldsymbol{s}^{\top}$, where $\boldsymbol{q} = \mathbf{X}\boldsymbol{s} / \boldsymbol{s}^{\top}\boldsymbol{s}$ 2） Collect $\boldsymbol{q}$ into $\mathbf{Q} = \left[\boldsymbol{q}_1 \quad \cdots \quad \boldsymbol{q}_k\right]^{\top}$ Return to Step 3 to extract the next SF until all k SFs have been obtained.
6.	SF extraction of historical observation $\mathbf{X}$ $\mathbf{S} = \mathbf{X}^{\top}\mathbf{P}^{\top}\left(\mathbf{Q}\mathbf{P}^{\top}\right)^{-1}$, $\boldsymbol{s}(t) = \left(\mathbf{P}\mathbf{Q}^{\top}\right)^{-1}\mathbf{P}\boldsymbol{x}(t)$
7.	One-step prediction given historical data $\bar{\mathbf{X}} = \left[\boldsymbol{x}(t) \quad \cdots \quad \boldsymbol{x}(t-l+1)\right]$ 1） $\hat{\boldsymbol{s}}(t+1) = \boldsymbol{\Lambda}_{\bar{\mathbf{S}}}\mathrm{vec}\left(\boldsymbol{\Phi}^{\top}\right)$, where $\boldsymbol{\Lambda}_{\bar{\mathbf{S}}} = \mathrm{diag}\left\{\left(\bar{\boldsymbol{s}}^{(1)}\right)^{\top} \quad \cdots \quad \left(\bar{\boldsymbol{s}}^{(k)}\right)^{\top}\right\}$, $\bar{\mathbf{S}} = \left[\bar{\boldsymbol{s}}^{(1)} \quad \cdots \quad \bar{\boldsymbol{s}}^{(k)}\right]^{\top} = \left(\bar{\mathbf{X}}^{\top}\mathbf{P}^{\top}\left(\mathbf{Q}\mathbf{P}^{\top}\right)^{-1}\right)^{\top}$ 2） $\hat{\boldsymbol{x}}(t+1) = \mathbf{Q}^{\top}\hat{\boldsymbol{s}}(t+1)$

3.2.5. Determining the hyperparameters

The model construction of LTSFA needs to determine two hyperparameters: the number of SF k and dynamic order l. In machine learning, the data set is always split into three subsets: training set, validation set, and test set. The model performance of each pair of hyperparameters is validated by the validation set after model training using the training set to select the optimal hyperparameters with the best performance.

For LTSFA, after all k SFs have been extracted, the residual $\mathbf{X}^{(k+1)}$ is expected to be white, and thus the autocorrelation and cross-correlation for nonzero lags will be approximately zero. The 95% confidence interval (CI) of the correlations (Shardt, 2015) will be used to verify model performance. The optimal hyperparameters correspond to the least violation for the 95% CI.

4. LTSFA-BASED PROCESS MONITORING

4.1. Observation-space partition

The residual after performing LTSFA on the observation is essentially white. However, it contains a static structure, which can be further explored using standard PCA. Hence, the observation space can be decoupled into a slow feature subspace, a static principal component subspace, and a residual subspace, that is,

$$\mathbf{X} = \mathbf{Q}^{\top}\mathbf{S}^{\top} + \mathbf{X}^{(k+1)} = \mathbf{Q}^{\top}\mathbf{S}^{\top} + \mathbf{W}\mathbf{T}^{\top} + \mathbf{E} \tag{33}$$

where $\mathbf{W} = \left[\boldsymbol{w}_1 \quad \cdots \quad \boldsymbol{w}_r\right]$ is the loading matrix and $\mathbf{T} = \left(\mathbf{X}^{(k+1)}\right)^{\top}\mathbf{W}$ is the score matrix of the principal components. The number of static principal components is determined by the method of the cumulative percentage of variance (CPV) (Valle, Li, & Qin, 1999).

4.2. Process monitoring indices design

After partitioning the space, we can monitor the variations in the subspaces to detect abnormal situations using Hotelling's T^2 and the squared prediction error (SPE), that is,

$$T_d^2 = \boldsymbol{s}^{\top}\boldsymbol{s} = \left(\mathbf{P}\mathbf{Q}^{\top}\right)^{-1}\mathbf{P}\boldsymbol{x}^{\top}\boldsymbol{x}\mathbf{P}^{\top}\left(\mathbf{Q}\mathbf{P}^{\top}\right)^{-1} \tag{34}$$

$$T_s^2 = \boldsymbol{t}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{t} = \left(\boldsymbol{x}^{(k+1)}\right)^{\top}\mathbf{W}\boldsymbol{\Omega}^{-1}\mathbf{W}^{\top}\boldsymbol{x}^{(k+1)} \tag{35}$$

$$SPE = \boldsymbol{e}^{\top}\boldsymbol{e} = \|\left(\mathbf{I} - \mathbf{W}\mathbf{W}^{\top}\right)\boldsymbol{x}^{(k+1)}\|^2 \tag{36}$$

where $\boldsymbol{\Omega} = \frac{1}{N}\mathbf{T}^{\top}\mathbf{T} = \mathrm{diag}\left\{\lambda_1,\cdots,\lambda_r\right\}$ and λ_i is the eigenvalue of $\mathbf{X}^{(k+1)}\left(\mathbf{X}^{(k+1)}\right)^{\top}$. The above three indices actually monitor the stationary variations in the subspaces. The dynamic behaviour anomaly can be detected by using the innovation as

$$T_v^2 = \boldsymbol{v}^{\top}\mathbf{P}_v\boldsymbol{\Lambda}_v^{-1}\mathbf{P}_v^{\top}\boldsymbol{v} \tag{37}$$

where $\boldsymbol{v}(t) = \boldsymbol{s}(t) - \hat{\boldsymbol{s}}(t)$ is the innovation and $\left\langle\boldsymbol{v}(t)\boldsymbol{v}^{\top}(t)\right\rangle = \mathbf{P}_v^{\top}\boldsymbol{\Lambda}_v\mathbf{P}_v$ is the singular value decomposition of the innovation covariance.

Under an assumption of a Gaussian distribution, the control limits for the indices can be determined as follow given a significance level α (Chiang, Russell, & Braatz, 2000; Qin, 2003)

$$T_d^2 \le T_d^2(\alpha) = \chi_k^2(\alpha) \tag{38}$$

$$T_s^2 \le T_s^2(\alpha) = \frac{r(N-1)(N+1)}{N(N-r)} F_{r,N-r}(\alpha) \tag{39}$$

$$SPE \le SPE(\alpha) = g\chi_h^2(\alpha) \tag{40}$$

$$T_v^2 \le \frac{k(N-1)(N+1)}{N(N-k)} F_{k,N-k}(\alpha) \tag{41}$$

where χ_k^2 and χ_h^2 are χ^2-distribution with k and h degrees of freedom, $F_{r,N-r}$ and $F_{k,N-k}$ are F-distributions with r and N-r, k and N-k degrees of freedom, $g = \theta_2 / \theta_1$, $h = \theta_1^2 / \theta_2$, and $\theta_j = \sum_{i=r+1}^{m} \lambda_i^j$ for $j = 1, 2$. If the indices of an observation exceed the control limits, then it is considered that a fault has occurred.

5. CASE STUDIES

5.1. Simulation example

A second-order autoregressive model is designed

$$\begin{cases} \boldsymbol{s}_t = \mathbf{A}_1 \boldsymbol{s}_{t-1} + \mathbf{A}_2 \boldsymbol{s}_{t-2} + \boldsymbol{v}_t \\ \boldsymbol{x}_t = \mathbf{B}\boldsymbol{s}_t + \boldsymbol{e}_t \end{cases} \tag{42}$$

$$\mathbf{A}_1 = \begin{bmatrix} 0.7904 & 0 & 0 \\ 0 & 0.5026 & 0 \\ 0 & 0 & 0.7904 \end{bmatrix}$$

$$\mathbf{A}_2 = \begin{bmatrix} 0.1004 & 0 & 0 \\ 0 & 0.3026 & 0 \\ 0 & 0 & -0.2104 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} -0.8207 & 0.5970 & 0.0253 \\ -0.8164 & -0.4752 & -0.6769 \\ -0.9511 & -0.5484 & 0.2549 \\ -0.3652 & -0.5081 & 0.4817 \\ 0.6245 & 0.7796 & 0.7808 \end{bmatrix}$$

where $\boldsymbol{e}_t \in \mathbb{R}^5 \sim N(0, 0.5^2\mathbf{I})$ and $\boldsymbol{v}_t \in \mathbb{R}^3 \sim N(0, 0.5^2\mathbf{I})$ are independent and identically distributed random processes.

Three thousand data samples are generated and split into training, validation, and test set with 2000, 500, and 500 data samples respectively. Using the method described in Section 3.2.5, the hyperparameters are determined as $l = 2$ and $k = 3$. The autocorrelation and cross-correlations of the original data $\boldsymbol{x}(t)$ and residual $\tilde{\boldsymbol{x}}(t)$ are shown in Figure 1 and Figure 2. It can be seen that the dynamics of the data, which is indicated by both autocorrelation and cross-correlation with nonzero delays, is filtered by LTSFA. Figure 3 is the residual correlations of the standard SFA. Assuming a maximum time delay of 20 samples, the percentage of times a point lies outside the 95% CI was 3.2% for LTSFA SFs, but 40.4% for the standard SFA. The correlations of the latent variables $\boldsymbol{s}(t)$ and the innovation $\boldsymbol{v}(t)$ in Figure 4 show that LTSFA can also uncover the dynamics of the latent states, while the standard SFA does not give an explicit expression of latent state dependency.

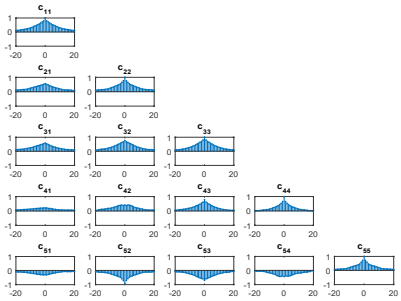

Figure 1: Autocorrelation and cross-correlation of $\boldsymbol{x}(t)$

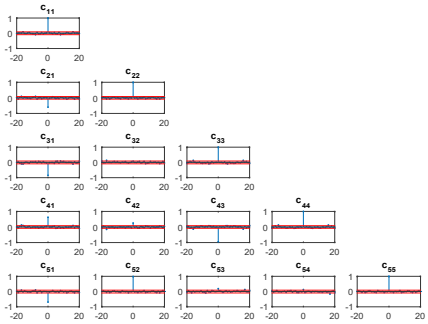

Figure 2: Autocorrelation and cross-correlation for the LTSFA residual

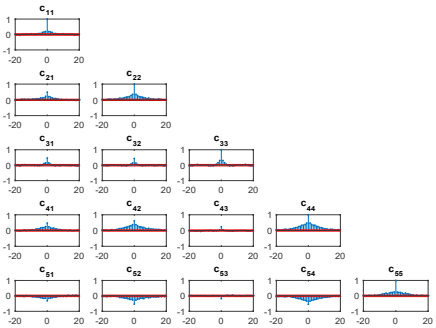

Figure 3: Autocorrelation and cross-correlation for the standard SFA residual

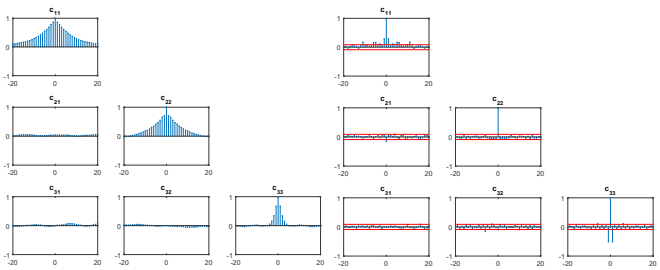

Figure 4: Correlations for left $\boldsymbol{s}(t)$ and right innovation $\boldsymbol{v}(t)$

5.2. Tennessee Eastman process benchmark

The Tennessee Eastman process (TEP) is a widely used benchmark in process control (Downs & Vogel, 1993). The revised version is adopted in this study (Lyman, Georgakis, & engineering, 1995). It contains 12 manipulated variables (XMV (1-12)) and 41 measurement variables which consist of 22 process variables (XMEAS (1-22)) and 19 quality variables (XMEAS (23-41)). In this paper, we use 11 manipulated variables (XMV (1-11)) and all 22 process variables. Five hundred normal samples sampled at 3-min intervals are used to train the models. The hyperparameters of LTSFA are determined as $l = 3$ and $k = 19$. The dynamic SFA provided in reference (Shang et al., 2015) is built to compare the monitoring results. The optimal hyperparameters of dynamic SFA are selected as $d = 2$ and $q = 0.55$. The monitoring results of LTSFA can give more detailed information of the process variations. For instance, Figure 5 shows that the variation of IDV(4) mainly occurs in the SF subspace and the residual subspace, while the static principal component subspace is less affected. Figure 6 shows that the anomaly caused by the high-frequency oscillation of IDV(14) can also be monitored in the SF subspace immediately. Furthermore, LTSFA only uses 19 SFs, while the dynamic SFA 46. Thus, the dynamic SFA may not be able to perform dimension reduction, especially in the case of large delays.

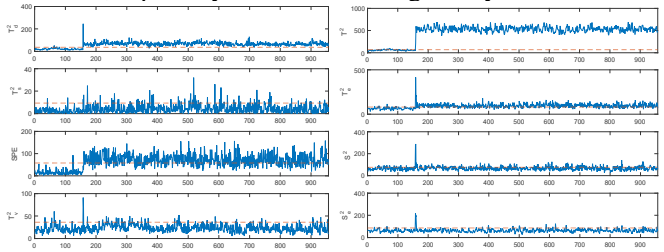

Figure 5: Monitoring results for IDV(4) using left LTSFA and right the dynamic SFA

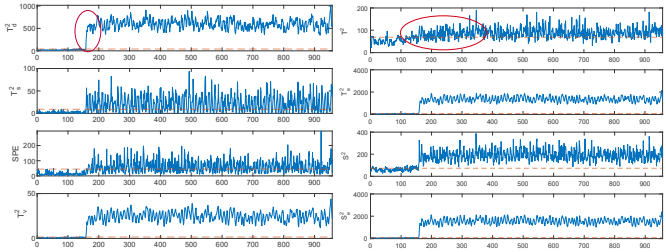

Figure 6: Monitoring results for IDV(14) using left LTSFA and right the dynamic SFA

6. CONCLUSIONS

This paper presented a new model called LTSFA to overcome the Markov assumption of the original SFA and improve the long-term dynamics modelling ability. The objective of LTSFA is established based on a high-order regressive model and an iterative algorithm is developed to solve the objective. In addition to the better dynamic modelling capability, LTSFA also gives an explicit expression of the long-term temporal dependency of latent states. A process monitoring strategy is developed using this approach. It is tested on a simulated example and the Tennessee Eastman process. It is shown that the proposed method can better extract the system dynamics and give more detailed process variations using fewer slow features than the standard SFA and the dynamic SFA methods. Further work will consider extending the results to nonlinear systems.

REFERENCES

Chiang, L. H., Russell, E. L., & Braatz, R. D. (2000). *Fault detection and diagnosis in industrial systems*: Springer Science & Business Media.

Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers chemical engineering, 17*(3), 245-255.

Gajjar, S., Kulahci, M., & Palazoglu, A. (2018). Real-time fault detection and diagnosis using sparse principal component analysis. *Journal of Process Control, 67*, 112-128.

Kano, M., Tanaka, S., Hasebe, S., Hashimoto, I., & Ohno, H. (2003). Monitoring independent components for fault detection. *AIChE Journal, 49*(4), 969-976.

Ku, W., Storer, R. H., & Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics intelligent laboratory systems, 30*(1), 179-196.

Lyman, P. R., Georgakis, C. J. C., & engineering, c. (1995). Plant-wide control of the Tennessee Eastman problem. *19*(3), 321-331.

Qin, S. J. (2003). Statistical process monitoring: basics and beyond. *Journal of Chemometrics: A Journal of the Chemometrics Society, 17*(8-9), 480-502.

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual reviews in control, 36*(2), 220-234.

Shang, C., Yang, F., Gao, X., Huang, X., Suykens, J. A., & Huang, D. (2015). Concurrent monitoring of operating condition deviations and process dynamics anomalies with slow feature analysis. *AIChE Journal, 61*(11), 3666-3682.

Shardt, Y. A.W. (2015). *Statistics for Chemical and Process Engineers*: Springer.

Shardt, Y. A.W., Zhao, Y., Qi, F., Lee, K., Yu, X., Huang, B., & Shah, S. (2012). Determining the state of a process control system: Current trends and future challenges. *The Canadian Journal of Chemical Engineering, 90*(2), 217-245.

Valle, S., Li, W., & Qin, S. J. (1999). Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Industrial Engineering Chemistry Research, 38*(11), 4389-4401.

Wise, B. M., Ricker, N., Veltkamp, D., & Kowalski, B. R. (1990). A theoretical basis for the use of principal component models for monitoring multivariate processes. *Process control quality, 1*(1), 41-51.

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural computation, 14*(4), 715-770.